systems

MDPI

Article

A Real Estate Price Index Forecasting Scheme Based on Online News Sentiment Analysis

Tao Xu [1,2], Yingying Zhao [1] and Jie Yu [3,*]

[1] School of Computer and Information Engineering, Henan University, Zhengzhou 450046, China; txu@henu.edu.cn (T.X.); zyy933@henu.edu.cn (Y.Z.)
[2] Henan Industrial Technology Academy of Spatio-Temporal Big Data, Henan University, Zhengzhou 450046, China
[3] School of College English Teaching and Research, Henan University, Zhengzhou 450046, China
* Correspondence: 10310101@vip.henu.edu.cn

Abstract: The real estate price index serves as a crucial indicator reflecting the operational status of the real estate market in China. However, it often lags until mid-next month, hindering stakeholders from grasping market trends in real time. Moreover, the real estate market has an extremely complex operating mechanism, which makes it difficult to accurately assess the impact of various policy and economic factors on the real estate price index. Therefore, we hope, from the perspective of data science, to explore the emotional fluctuations of the public towards the real estate market and to reveal the dynamic relationship between the real estate price index and online news sentiment. Leveraging massive online news data, we propose a forecasting scheme for the real estate price index that abandons complex policy and economic data dependence and is solely based on common and easily obtainable online news data. This scheme involves crawling historical online real estate news data in China, employing a BERT-based sentiment analysis model to identify news sentiment, and subsequently aggregating the monthly Real Estate Sentiment (RES) index for Chinese cities. Furthermore, we construct a Vector Autoregression (VAR) model using the historical RES index and housing price index to forecast future housing price indices. Extensive empirical research has been conducted in Beijing, Shanghai, Guangzhou, and Shenzhen, China, to explore the dynamic interaction between the RES index and both the new housing price index and the second-hand housing price index. Experimental results showcase the unique features of the proposed RES index in various cities and demonstrate the effectiveness and utility of our proposed forecasting scheme for the real estate price index.

Keywords: real estate price index; online news; sentiment analysis; BERT; VAR; time series forecasting

Academic Editor: Vladimír Bureš

Received: 27 November 2024
Revised: 4 January 2025
Accepted: 7 January 2025
Published: 8 January 2025

Citation: Xu, T.; Zhao, Y.; Yu, J. A Real Estate Price Index Forecasting Scheme Based on Online News Sentiment Analysis. *Systems* **2025**, *13*, 42. https://doi.org/10.3390/systems13010042

1. Introduction

The real estate market is a dynamic and complex system, affected by the financial market, economic policies, land prices, emergencies, mortgage rates, and other factors [1]. Because most Chinese people regard housing as a necessity for marriage and children's education [2], house prices have become the people's and policymakers' key concern. Therefore, an accurate forecasting tool for the real estate market is not only useful for the formulation of scientific policies but also has great guiding significance.

As a key accessing metric of the real estate market, the real estate price index offers an insightful representation of the market's operational status. The general practice in many countries is that the national authoritative department regularly publishes this

index, providing stakeholders with critical insights into the evolving trends of the market. However, owing to the complexity of the economic statistics workload, the release of the real estate price index always has significant lags. For example, the monthly real estate price index of 70 cities in China released by the National Bureau of Statistics of China usually has a delay period of more than 15 days, that is, 50% of the statistical period [3]. This imposes limitations on objectively and timely evaluating the dynamics of the real estate market [4]. Specifically, the delay is a significant information friction with measurable effects on important economic variables. For example, a release of the house price index has an immediate effect on the stock prices of home-building companies, despite the fact that this release contains information about housing market conditions from a few months earlier. If the stock market is unable to overcome the reporting delays associated with house prices, it seems likely that individual homeowners, policymakers, lenders, etc., are not either, suggesting that this information friction may have much broader effects on financial markets and real economic activity. Moreover, for market participants who urgently need to understand the trends in housing prices, its impact cannot be overlooked. For example, for ordinary home buyers, it may be difficult for them to grasp housing prices in a timely manner due to information lag, and their housing purchase costs may increase. For real estate enterprises, a 15-day lag may lead to a situation where their strategy formulation is influenced by asymmetric and lagging indices, interfering with the operation and development of the enterprises. Financial institutions rely on timely housing price information when evaluating real estate. A 15-day lag will increase financial risks and affect market stability.

Furthermore, due to the unique nature of the real estate market and the diversity of statistical data sources, countries adopt diverse methods for price index construction, resulting in the absence of a uniform standard. Previous studies mainly focus on the impact of economic factors on housing prices, such as macroeconomic indicators (GDP growth, interest rate level, inflation rate, etc.) [5–7], the fundamental aspects of supply and demand in the real estate market (population changes, land supply, housing properties, number of transactions, etc.) [8–12], and policy regulation factors (purchase restrictions, loan restrictions, tax policies, etc.) [11,13–16]. In addition, under the guidance of the law of market economy, the mainstream methods of compiling real estate price indices are characteristic price method [17–19], repeat sale method [20–22], and pooled method [10,23]. These methods, rooted in the extensive collection of economic data, employ statistical methods such as weighted average [24], regression [25], generalized least squares [26,27], and maximum likelihood estimation [28] to estimate the real estate price index. Their effectiveness and robustness, particularly the accuracy in describing market dynamics, remain to be thoroughly verified [17,29,30].

Meanwhile, big data technology has enhanced the forecasting performance of various economic industries, such as real estate, oil, and stocks. Currently, economic forecasting methods can be divided into two categories: web search data (WSD) and online textual data. The main idea of the WSD category is to collect the Google Trends and Baidu Index as exogenous variables to improve forecasting accuracy [2]. Wang et al. [31] proposed a new framework utilizing WSD for crude oil price forecasting, suggesting that the Baidu Index is stronger than Google Trends in terms of predictive power in the Chinese context. Similarly, Coble et al. [32] used Google Trends to predict building permits, and they showed that models including Google search queries nowcast and forecast better than many of our good (but not naive) benchmarks. With the development of news media platforms, a large number of online news texts have been generated, which can provide effective and convincing information. Considering the real estate market as an example, by delving into the vast and diverse news content available on the Internet, researchers can capture

comprehensive public sentiment regarding market fluctuations [33–35] and, importantly, gain insights [36] into the evolving trends of the real estate market without delay.

In order to grasp the real estate market dynamics for the first time, some studies have been conducted to explore the relationships between online news and the real estate market. Kang et al. [37] proposed a short-term forecasting model of apartment prices based on the news search popularity. Nowak et al. [38] proposed a regularized regression model-based method to identify real estate-related words in the news. Isler et al. [39] experimentally studied bounded rationality in real estate by observing the effects of market news and media credibility cues on house price forecasting. Focusing on online news sentiment, Hausler et al. [40] examined the relationship between news-based sentiment, captured through a machine learning approach, and the US-securitized and direct commercial real estate markets. Soo et al. [41] quantified local housing news articles to construct the housing sentiment index (SI), and they verified that the housing SI has excellent predictive power for future house price forecasting. Beracha et al. [42] demonstrated that news-based sentiment could be used as an early market indicator in the United States. McCarthy and Alaghband [43] demonstrated a modest correlation between sentiment and market movements. Kim et al. [44] utilized sentiment analysis on news articles to generate a News Sentiment Index score, which is then seamlessly integrated into a real estate price forecasting model. Shao et al. [45] used wavelet analysis to explore the dynamic relationship between house prices and online news sentiment in China and found that emotion also had a significant impact on house prices in third-tier cities and western regions of China. However, due to the difficulty of quantifying sentiment and the lack of data, only a few studies have applied sentiment to the Chinese real estate market. These studies employ machine learning methods to explore the quantitative relationship between news sentiment and the real estate market. However, due to the complexity of unstructured text data and the limitations of machine learning methods, they cannot accurately interpret massive news data, resulting in low accuracy of experimental results [46] and limited application scenarios [47].

Real estate online news not only reports recent facts and events but also reflects the author's preferences and emotional tendencies toward them. Identifying the sentiments expressed in news articles is crucial for exploring dynamic trends in the real estate market. Sentiment analysis is an important task in Natural Language Processing (NLP), and there are already several pre-trained models based on massive textual data used to solve it, such as BERT (Bidirectional Encoder Representation from Transformers) and GPT (Generative Pre-Trained Transformer). BERT uses multiple bidirectional Transformer structures to simultaneously learn contextual information in text, making it suitable for various natural language understanding (NLU) tasks; GPT uses a unidirectional (left to right) Transformer model, which excels at generating coherent text and is more suitable for natural language generation (NLG) tasks. Some recent researches [48–51] have proposed some BERT-based schema for news sentiment analysis and mined the quantitative relationships between news sentiment and specific real estate indicators, such as sale price [52], rent price [53], and stock price [54]. Therefore, the BERT model is more suitable for learning the contextual information of online news and analyzing news sentiment. However, due to the fact that the fine-tuned Bert model with different news datasets is not universal [55], it is necessary to design a special scheme to deal with the news sentiment analysis in a specific real estate market.

Motivated by the above analysis, we introduce a novel approach based on sentiment analysis of online news to evaluate and predict real estate price index, aiming to study the real estate price prediction problem from the perspective of data science, avoid complex market policies and economic factors, and focus only on exploring the correlation between public sentiment and real estate prices index. Firstly, we crawl massive real estate online

news data of specific cities from Baidu News (news.baidu.com), the most influential Chinese Internet search engine; we preprocess and structure these news texts to obtain a news' dataset that can comprehensively describe the historical dynamics of the real estate market. Secondly, we implement the fine-tuning of the free pre-trained BERT model on the real estate news dataset and quantify the sentiment of each news article. Thirdly, we aggregate the news sentiment values monthly to get the monthly Real Estate Sentiment (named RES) index of a city, and we also collect two types of real estate price indices for specific Chinese cities, i.e., the monthly new housing price index and the monthly second-hand housing price index, from National Bureau of Statistics of China. Then we employ the classical VAR (Vector Autoregression) method [56,57] to build a forecasting model for the RES index and the monthly real estate price index. Finally, we achieve the non-delayed forecasting for the monthly housing price index.

To verify the effectiveness of the proposed scheme, we conducted sufficient empirical research in four super first-tier cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen. Experimental results show that, in these four cities, there are high correlations between the RES index and the monthly real estate price index compiled by the National Bureau of Statistics of China. This not only demonstrates that we can use online news information to forecast the real estate price index without delay but also provides some intuitive evidence for the rationality of the proposed RES index.

The main contributions of this paper are summarized as follows:

- We construct a Chinese real estate online news dataset, which includes massive online news about real estate from 2011 to 2022 in four Chinese cities: Beijing, Shanghai, Guangzhou, and Shenzhen. Based on them, we built the RES index for each city.
- We introduce a novel real estate price index forecasting scheme based on sentiment analysis of online news. This scheme fine-tunes a pre-training BERT model to conduct accurate sentiment analysis on the Chinese real estate online news dataset and generates the RES index for Chinese cities. Then it constructs the VAR model between the RES index and the real estate price index to achieve an accurate evaluation of the monthly real estate price index.
- Empirical research results from four Chinese cities, Beijing, Shanghai, Guangzhou, and Shenzhen, show that there are strong correlations between online news information and the real estate price index compiled by the National Bureau of Statistics of China. These results also demonstrate that the proposed forecasting scheme can accurately forecast the monthly real estate price index without delay using the current month's real estate information data.

The subsequent sections of this paper are outlined as follows: Section 2 defines the proposed scheme and algorithm, Section 3 introduces the experimental findings, and Section 4 presents the conclusion of the paper, respectively.

2. Methodology

As shown in Figure 1, the proposed real estate price index forecasting scheme consists of 4 modules. In the News acquisition and preprocessing module, we crawl real estate online news for specific cities from the Internet and structure them into a real estate news dataset. In the News sentiment analysis module, we employ the pre-trained BERT model to fine-tune the features of news texts and understand the sentiment expressed in the news. In the Modeling RES index stage, we aggregate the sentiment values of each news record into a monthly RES index series. Then, we use the VAR method to create a forecasting model that can capture the interaction mechanism between the RES index and the price index

published by the National Bureau of Statistics of China. In the Forecasting stage, we use this forecasting model to forecast the next monthly real estate price index and RES index.

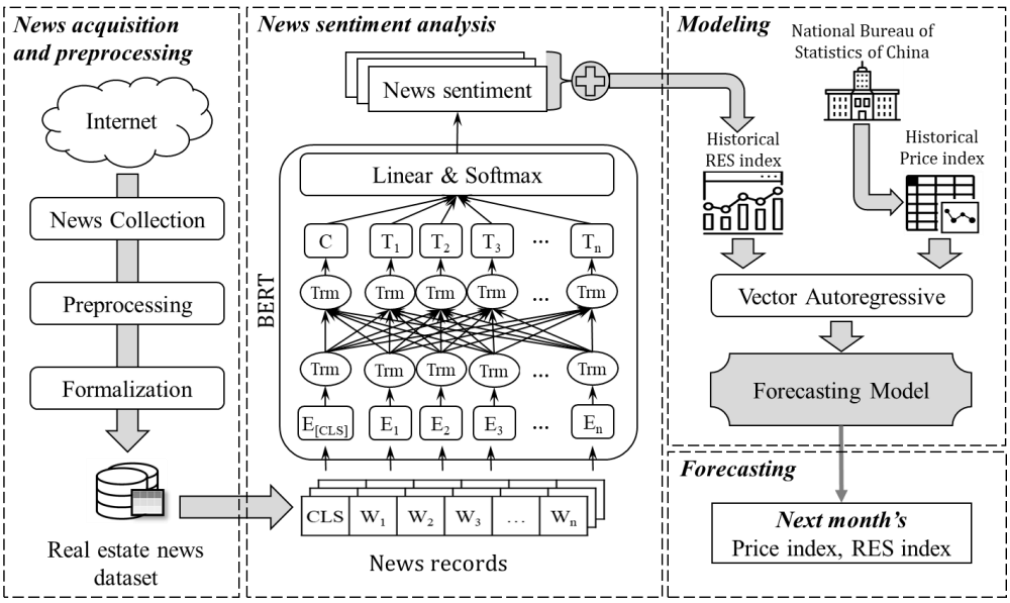

Figure 1. The overview of the proposed scheme.

2.1. News Acquisition and Preprocessing

Our goal is to study the real estate market in the urban area. Firstly, we crawl real estate online news from January 2011 to December 2022 for these four cities in China from the website, news.baidu.com, and download and organize monthly real estate price index series data for new houses and second-hand houses from the National Bureau of Statistics of China, stats.gov.cn. Then we preprocess the downloaded news data by data augmentation, data deduplication, and emotional annotation. According to the tendency to affect housing prices, news emotions are divided into positive, medium, and negative categories and marked manually. Finally, the data are formatted. The specific steps are described in Section 3.1.

Let $\mathbb{D} = \{D^{city}\}$ be the preprocessed real estate online news dataset, where D^{city} is a subset of online news for a city. Let $d_i^{city} = \{X,\ date, city, label\}$ be the i-th news record in D^{city}, $d_i^{city} \in D^{city}$, where X is the news text, $date$ is the publication date of the news, and $label \in [1, 0, -1]$ represents the 3 sentiment levels (i.e., "positive", "neutral", and "negative") of news texts based on their tendency towards the real estate market.

Additionally, we also collect monthly series data on the new housing price index and second-hand housing price index from the National Bureau of Statistics of China (stats.gov.cn) for these specific cities during the same period. Let $nP^{city} = \{np_{date}\}$ and $sP^{city} = \{sp_{date}\}$ be the time series for the new housing price index and second-hand housing price index, where np_{date} and sp_{date}, respectively, represent the new housing price index and the second-hand housing price index for a given $date$.

Based on the above definitions, we aim to explore the interaction patterns between the two real estate price indices and the sentiment expressed in online news.

2.2. News Sentiment Analysis

To clarify the tendency of online news towards the development trends of the real estate market, we develop a BERT-based framework to identify the sentiment of news text. As shown in Figure 2, we first format the news text into embedding representations, then use a pre-trained bidirectional Transformer structure to learn the features of news text, and

finally design a Linear layer and a Softmax layer to output the probability that the news belongs to each sentiment category. The sentiment type with the highest probability value is the news sentiment.

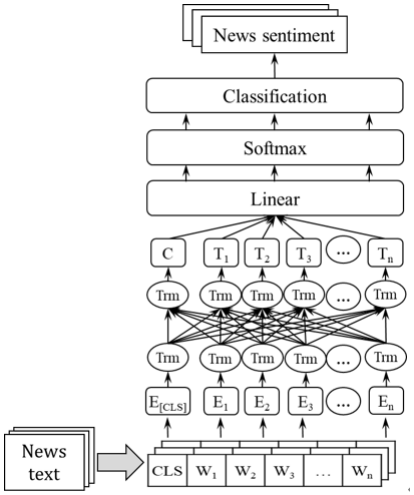

Figure 2. The BERT-based news sentiment analysis model.

The formatting steps of the original news text include tokenization and embeddings. We use the WordPiece algorithm [58] to tokenize Chinese news text X into word vectors $W =< \text{CLS}, w_1, w_2, \ldots, \text{SEP}, w_i, \ldots, \text{SEP} >$, where w_i is the i-th word, "CLS" is a beginning token of a sentence, and "SEP" is a clause token. To fully understand the semantics of the text, BERT uses a combination of three embeddings of tokens as input, i.e., word embeddings, sentence embeddings, and positional embeddings. We denote E_i as the comprehensive embedding of the i-th token and $E =< E_i >$ as the input embedding vector of BERT.

The process of feature learning from news text involves two stages: pre-training and fine-tuning, both of which optimize model parameters with the same bidirectional Transformer structure. As shown in Figure 2, the Transformer unit is represented as Trm, and the learned hidden embeddings are represented as $T =< T_i >$.

In the pre-training stage, the model learns the bidirectional contextual relationships of vocabulary on large-scale unlabeled data and understands the logical relationships between sentences. In this paper, we employ a pre-trained Chinese BERT model for Chinese text learning.

In the fine-tuning stage, the model fine-tunes the pre-trained parameters on the labeled news data for news sentiment analysis. The linear layer is responsible for transforming high-dimensional embedding results into the category dimension of news semantics, i.e., 3. The transformation method is shown in Equation (1):

$$O = W \cdot T + b, \tag{1}$$

where O represents the output of the linear layer, W denotes the weights, and b stands for the biases.

The Softmax layer normalizes O to express the probability of the sentiment category for the input text. We employ the cross-entropy loss [59] as the loss function in the fine-tuning stage, as shown in Equation (2):

$$L = \sum_{i=1}^{3} (-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)), \tag{2}$$

where L represents the loss value, y_i is the predicted value, and $\hat{y}_i$ is the real value.

In the fine-tuning process, we first split the news data into a training set and a testing set. Then, we fine-tune the model on the training set with the aim of minimizing cross-entropy loss. During this process, we adjust the training parameters, such as batch size, learning rate, and the number of training epochs. Finally, we utilize the trained model to analyze the news data in the testing set, determine the sentiment category label with the highest probability in the classification layer, and output the sentiment value of the news data.

2.3. Modeling RES Index

The RES index is defined as the average sentiment of online news in a city over a period of time, which expresses the public opinions towards the real estate market. Furthermore, we employ the VAR model [57] to quantitatively evaluate both the RES index series and the real estate price index series and examine the long-term stable relationship between them.

Let $R = \{r_i | i \geq 0\}$ be the RES index series, r_i be the ith RES index. As shown in Equation (3), the RES index r is the average sentiment value of online news during a specific time period, i.e., $d.date \in r.timespan$.

$$r = \frac{1}{n}\sum d.label,\ n > 0\ and\ d.date \in r.timespan, \tag{3}$$

where $d.label$ represents the sentiment value of a news record d. If $n = 0$, i.e., there is no news data within $r.timespan$, we use the linear interpolation method to fill in r.

To obtain the sentiment value of news records, i.e., $d.label$, we train the BERT model described in Section 2.2 using data outside the target period as the training set and predict the sentiment value of the news during the target period. We then average the predicted values to generate the RES index for that period. The process of solving the RES index for a given period is shown in Algorithm 1.

Algorithm 1: Solving the RES index for a given period.

Input: D: the real estate news dataset, $\mathbb{B}$: the pre-trained BERT model, T: the given period.
Output: r: the RES index for period T.

$trainSet = \{\ \}$; $testSet = \{\ \}$;
for each $d \in D$:
    if $d.date \in T$:
        $testSet = d.X \cup testSet$;
    else:
        $trainSet = [d.X, d.label] \cup trainSet$;
$\mathbb{B}^{'} \leftarrow \mathbb{B}(trainSet)$ //fine-tune, $\mathbb{B}^{'}$ is the trained BERT model.
$L \leftarrow \mathbb{B}^{'}(testSet)$ //predict, L is the set of news sentiments in $testSet$.
$r \leftarrow Equation\ 3(L)$ //solve the RES index.
return r

2.4. Constructing VAR Forecasting Model

The VAR model (Vector Autoregressive model), proposed by Christopher Sims [57], is a widely popular method for multivariate time series analysis, enabling the evaluation of dynamic relationships among multiple variables. It constructs models solely based on the statistical properties of the data, without any reliance on preconceived constraint conditions, and is frequently employed for dynamic modeling and forecasting of multiple interrelated economic indicators. In this paper, we incorporate the RES index series obtained through news sentiment analysis and the real estate price into the VAR model. By analyzing the historical data of the two variables, the VAR model can estimate the coefficients that describe the correlation between each variable (i.e., RES index and real estate price index) and its own past values as well as the past values of the other variable. In the forecasting

stage, with the estimated VAR model and historical data, the model will calculate the forecasted values for the next period of the real estate price index and the RES index based on its internal mechanism. The specific process is as follows:

Let $C = \{c_i | i \geq 0\}cc$ be the real estate price index time series. The k-order VAR model for both R and C is as shown in Equation (4):

$$\hat{Y}_t = A_0 + A_1\hat{Y}_{t-1} + A_2\hat{Y}_{t-2} + \ldots + A_k\hat{Y}_{t-k} + \varepsilon_t, \tag{4}$$

where $\hat{Y}_t = [c_t, h_t]'$ is a vector of the real estate prices index and the RES index for time span t, $A_0 = [a_{c0}, a_{h0}]'$ is a constant vector, $A_i = \begin{bmatrix} a_{c1,i} & a_{c2,i} \\ a_{h1,i} & a_{h2,i} \end{bmatrix}$ is a parameters matrix, $k \geq i > 1$, k is the lag order, and $\varepsilon_t = [\varepsilon_{ct}, \varepsilon_{ht}]'$ is a residual vector.

Most applications of the Vector Autoregression (VAR) model are based on the assumption of stationary data. Stationary data help to ensure the consistency and reliability of model parameter estimation and, at the same time, can avoid the spurious regression problem caused by non-stationary data. Although the cointegrated VAR can directly handle non-stationary data, it has limitations. For example, the test for cointegration relationships is complex and the results are uncertain, and it is not as intuitive and effective as the VAR model based on stationary data in explaining short-term dynamic adjustments. Therefore, VAR models be constructed on stationary time series, meaning that the modulus of the model's eigenvalues must be less than 1, or equivalently, the eigenvalues must lie within the unit circle. We utilize the Augmented Dickey–Fuller (ADF) method [60] to test the stationarity of R and C. If any of these variables are found to be non-stationary, both must undergo differencing until they both achieve stationarity, which can be represented by Equation (5).

$$\hat{Y}_t = \begin{cases} \hat{Y}_t, & R \text{ and } C \text{ are both stationary time series} \\ \hat{Y}_t - \hat{Y}_{t-1}, & else \end{cases}, \tag{5}$$

Next, we also need to solve for the optimal lag order k of the VAR model. We employ the Akaike Information Criterion (AIC) indicator [61], which is a measure of the goodness of fit of a statistical model and indicates a better model with a smaller value. The AIC of the VAR model with k lag periods can be calculated using Equation (6):

$$AIC(k) = -2 * \ln(L) + 2 * k, \tag{6}$$

where k is the number of lag periods, and L is the likelihood value of the VAR model. We traverse various values of k and calculate the $AIC(k)$ for the VAR model. We iterate through the k value to compute the $AIC(k)$ of the VAR model, and the optimal value of lag periods k is determined by selecting the value that yields the minimum AIC.

Finally, based on the lag order k, we need to estimate the parameters of the VAR model, i.e., A_0, A_i, and ε_t in Equation (4). Since C and R are stationary time series, we can assume that their mean and variance are close to 0. We use the classical OLS method to estimate the parameters of the VAR model. Although the OLS method has certain biases in the context of time series, it possesses the property of consistency. As numerous studies have indicated [62,63], when dealing with similar stationary time series data, OLS can gradually converge to the true parameter values with the increase in data volume, providing reliable results for solving the VAR parameters. The parameters can be solved by Equation (7):

$$A = \left(Y^T Y\right)^{-1} Y^T \hat{Y}_t, \tag{7}$$

where $A = [A_0, A_1, \ldots, A_k]$ is a parameter matrix set, and $Y = [Y_{t-1}, \ldots, Y_{t-k}]$ is an argument matrix. After obtaining A, input A into the VAR model to solve for the residual vector ε_t, as defined in Equation (4).

Based on the above, the VAR model for the RES index and the real estate price index has been successfully constructed. The modeling process can be formalized as Algorithm 2.

Algorithm 2: Modeling the VAR model for the RES index and the real estate price index.

Input: D: the real estate news dataset, $\mathbb{B}$: the pre-trained BERT model, T: the given period.
Output: M: the constructed VAR model.

$R \leftarrow Algorithm1(D, \mathbb{B}, T)$; //The RES index time series
while $ADF(R)$ or $ADF(C)$ is not stationary: //Equation 6
    $R, C \leftarrow Equation\ 5(R, C)$; // Time series differencing
for i in $[1,\ n]$:
    if $AIC(i)$ is minimum:
        $k \leftarrow i$; //the lag period
$A \leftarrow Equation\ 7(k, R, C)$; //VAR parameter estimation
$M \leftarrow Equation\ 4(k, A)$
return M

The inputs of Algorithm 2 are the historical RES index time series R and the historical real estate price index time series C. Its outputs are the future RES index at time t, r_t^0, and the future real estate price index at time t, c_t^0. Algorithm 2 directly calls Algorithm 1 to construct the VAR model M and then determines whether the result is a differential value. If so, it is restored to its original value using Equation (5), and the predicted future r_t^0 and c_t^0 are output.

Based on the constructed VAR model, we can use historical data from the first k periods to predict the RES index and the real estate price index at time t. The forecasting process is shown in Algorithm 3.

Algorithm 3. Forecasting the future RES index and real estate price index.

Input: R: the historical RES index time series, C: the historical real estate price index time series.
Output: r_t^0: the future RES index at time t, c_t^0: the future real estate price index at time t.

$M \leftarrow Algorithm\ 2(R, C)$;
if $M.\hat{Y}_t$ is a differential value:
    $r_t^0, c_t^0 = Equation\ 4(M.\hat{Y}_t)$;
return r_t^0, c_t^0

3. Empirical Study and Result Discussion on China's Urban Real Estate Price Index

In this section, we take four super first-tier cities in China, i.e., Beijing, Shanghai, Guangzhou, and Shenzhen, as research objects, identify quantitative correlation patterns between news data and the real estate market, and aim to forecast future real estate price indices.

3.1. Acquisition and Preprocessing of Real Estate Data for Constructing VAR Forecasting Model

Data acquisition. We crawl real estate online news from January 2011 to December 2022 for these four cities in China from the website news.baidu.com (accessed on 21 March 2023). The Baidu.com website is the most popular Chinese Internet search engine, which can provide mainstream, instant, and massive aggregation of online news information.

We utilized "city, date, content keyword" as search keywords to crawl online news data. The targets of "city" are Beijing, Shanghai, Guangzhou, and Shenzhen. The unit of collection date is month, from January 2011 to December 2022. And the "content keywords"

are "housing price", "real estate", or "real estate market". By traversing all keyword combinations, we crawled 38,763 online news records, including 7763 in Beijing, 11,919 in Shanghai, 7778 in Guangzhou, and 11,303 in Shenzhen. Each record contains four attributes: headline, news body, date, and city.

Additionally, we downloaded and organized monthly real estate price index series data for new houses and second-hand houses from the National Bureau of Statistics of China, using them to explore the dynamic changes in the real estate market.

News data preprocessing. The preprocessing of online news includes the following steps:

(1) Data augmentation. Since some news writers often prefer exaggerated headlines to enhance the appeal of news stories, this often results in a mismatch in sentiment between the news headline and the actual content. Consequently, to accurately grasp the sentiments in the news text, we segregate the headline and content of a news article into two data records, which can not only clarify the sentiment of the text but also provide more data for news sentiment analysis.

(2) Data deduplication. To avoid the interference of duplicate news on sentiment analysis, we need to perform news record deduplicate based on news text. After data enhancement and deduplication, there are 30,522 news data in total, including 7018 in Beijing, 10,730 in Shanghai, 6868 in Guangzhou, and 2953 in Shenzhen.

(3) Sentiment annotation. To create a learnable news dataset, we categorize the sentiment of news into positive, neutral, or negative, based on its propensity to impact real estate price, and manually label each news record accordingly. We label news of rising real estate prices as positive, news describing a decline in real estate prices as negative, and news without obvious or opposing tendencies as neutral.

Data formalization. As defined in Section 2.3, $\mathbb{D} = \{D^{city}\}$ is the preprocessed real estate news dataset, where $city \in \{$"Bejing", "Shanghai", "Guangzhou", "Shenzhen"$\}$ is the super first-tier city in China. $d_i^{city} = \{X, date, city, label\}$ is the ith news record in D^{city}, where $d_i^{city} \in \mathbb{D}$, X is the news text, $date \in [Jan.2011, Dec.2022]$ is the publication date of the news, and $label \in [1, 0, -1]$ represents the three sentiment levels (i.e., "positive", "neutral", and "negative") of news texts based on their tendency towards the real estate market. Table 1 shows the statistics of the real estate news dataset $\mathbb{D}$. Shanghai has the largest number of news, followed by Guangzhou and Shenzhen, while Beijing has the lowest. This indicates the differences in attention to the real estate markets in various cities. Similarly, the total number of words is also the same trend.

Table 1. The statistics of the real estate news dataset.

Dataset	$D^{Beijing}$	$D^{Shanghai}$	$D^{Guangzhou}$	$D^{Shenzhen}$	$\mathbb{D}$ (Total)
The number of news	7018	10,730	6868	5906	30,522
The number of words	469,197	692,227	454,961	390,238	2,006,623

Based on the definition of the real estate price index in Section 2.1, we denote $nP^{Beijing}$, $nP^{Shanghai}$, $nP^{Guangzhou}$, and $nP^{Shenzhen}$ be the new housing price index series of four cities, and $sP^{Beijing}$, $sP^{Shanghai}$, $sP^{Guangzhou}$, and $sP^{Shenzhen}$ be their second-hand housing price index series. These data are normalized month-on-month time series from January 2011 to December 2022. We download the new housing price index and second-hand price index of four cities from the National Bureau of Statistics of China.

3.2. BERT-Based News Sentiment Analysis

To quickly obtain accurate semantic information on massive real estate online news, we introduce a BERT-based news sentiment analysis scheme. We downloaded the open-source BERT model pre-trained in Chinese text and fine-tuned it on the real estate news dataset.

Experimental settings. The open-source Chinese pre-trained BERT model is called "BERT-Base-Chinese", downloaded from https://huggingface.co/ (accessed on 1 April 2023). It adopts the basic BERT architecture, consisting of 12 layers of transformer encoders, each equipped with 12 self-attention heads. It has a total of 110 million parameters and was pre-trained using approximately 250 million words from the Chinese Wikipedia.

In the fine-tuning stage, the news dataset's training, validation, and test set proportions are 0.6, 0.2, and 0.2, respectively. We set the length of the news text input to the pre-trained BERT model to 128 and fill in any gaps with a zero vector. We use the Adam optimizer [64] for training with an initial learning rate of 2×10^{-5}. We experimented with the epoch size [3,5,7] and the batch size 32 to select the best parameters. We conducted all experiments in a TensorFlow environment with an NVIDIA 4060Ti GPU (32 GB memory).

Evaluation Metrics. To evaluate the performance of the BERT model, we employ *Precision*, *Recall*, and *F*1 (F1 score) as evaluation metrics. Higher *Precision*, *Recall*, and *F*1 values indicate better model performance. The formal definitions of them are as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN}, \tag{9}$$

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{10}$$

where TP denotes the number of correctly predicted positive samples, and TN denotes the number of correctly predicted negative samples. On the other hand, FP corresponds to the number of falsely predicted positive samples, and FN represents the number of falsely predicted negative samples.

Experimental results. The experimental results of fine-tuning the BERT model for sentiment analysis are shown in Table 2. It is worth noting that we achieve satisfactory performance on all real estate news datasets, with *precision*, *recall*, and *F*1 score consistently above 80%. This suggests that fine-tuning the BERT model to handle the real estate sentiment dataset is quite successful. We observe that $D^{Guangzhou}$ and $D^{Shenzhen}$ perform well on all metrics, while $D^{Beijing}$ performs relatively poorly. That is because the quality of the data also plays a role in the model forecasting, and different data sets may produce different results. Each city has its own unique language style and cultural characteristics that affect the way emotions are expressed. Overall, the performance of the BERT model is relatively consistent, indicating that the model has good generalization ability for sentiment analysis tasks in Chinese real estate news datasets.

Table 2. Sentiment analysis experiment results on real estate online news datasets.

Metrics	*Precision*	*Recall*	*F*1
$D^{Beijing}$	80.65	80.22	80.30
$D^{Shanghai}$	82.76	82.66	82.68
$D^{Guangzhou}$	85.64	84.59	84.73
$D^{Shenzhen}$	83.66	83.54	83.52

3.3. Building China's Urban RES Index

To obtain the RES index series of each city, we conduct Algorithm 1 (described in Section 2.3) to aggregate the sentiment values of real estate online news predicted by the BERT model on a monthly basis as the RES index for that month. The RES index of Beijing, Shanghai, Guangzhou, and Shenzhen from 2011 to 2022 is shown in Figure 3.

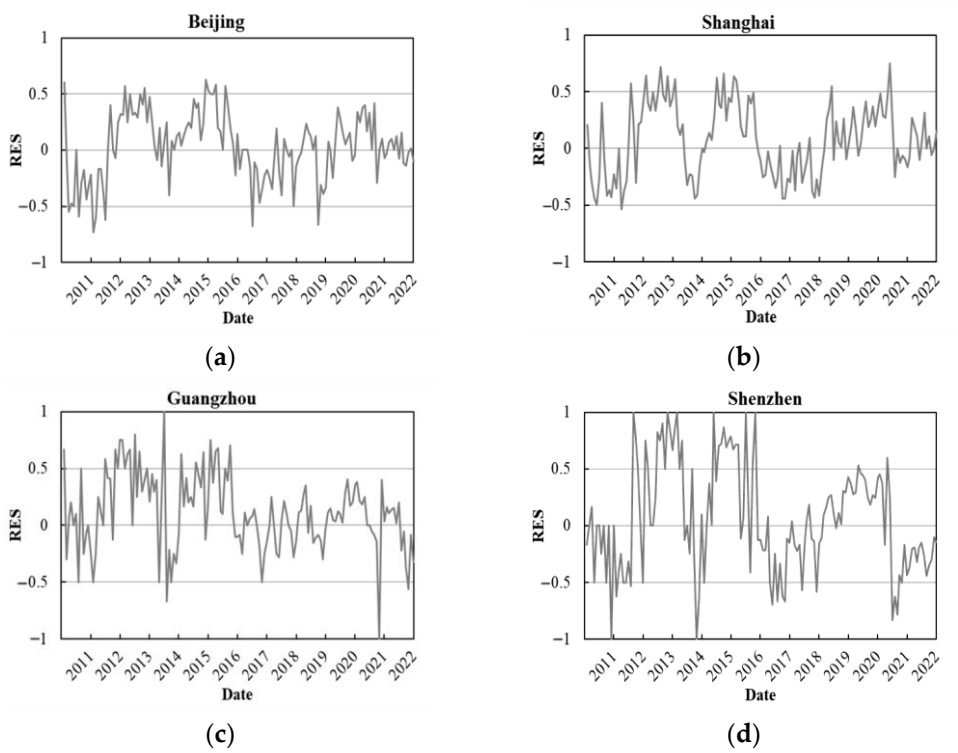

Figure 3. The RES index in four Chinese cities from 2011 to 2022: (**a**) $R^{Beijing}$, (**b**) $R^{Shanghai}$, (**c**) $R^{Guangzhou}$, and (**d**) $R^{Shenzhen}$.

As shown in Figure 3, the trends of the RES index series exhibit chaos and disorder, lacking obvious periodicity. This is due to the fact that the sentiment of online news is influenced by multiple factors, such as market supply and demand, economic situations, and national policies, with the underlying mechanism of change being intricately complex. The RES index series showed two notable peaks during the periods 2013–2014 and 2016–2017, reflecting a relatively optimistic sentiment among the public towards the real estate market. It is worth noting that during the special period of strict policy regulation in the third quarter of 2016 and when the market entered a depression stage in the second quarter of 2020, the online news sentiment index was able to sensitively respond with corresponding changes to the fluctuations in housing prices. Especially when the COVID-19 pandemic in 2020, as a sudden external shock, had a severe impact on the real estate market, the RES index also exhibited significant fluctuations. Overall, within the sample interval, the time series related to the RES index and the real estate market was clearly affected by both changes in the economic environment and adjustments in regulatory policies, thus presenting relatively obvious fluctuation characteristics, fully demonstrating its sensitivity and relevance to market dynamics.

Table 3 shows the statistical features of the RES index in four Chinese cities. We can observe that the highest Variance and Standard Deviation are in $R^{Shenzhen}$, while the other three cities are lower. This indicates that the fluctuation of $R^{Shenzhen}$ is more severe, as can also be seen from Figure 3. In terms of the Average and Median, the two values of $R^{Guangzhou}$ are 0.116 and 0.112, which are higher than those of other cities, indicating that the real estate market in Guangzhou is more optimistic. Moreover, the difference between Average and Median is the largest in $R^{Shenzhen}$, and the Average is positive, reflecting that $R^{Shenzhen}$ has more optimistic extreme values. In addition, the maximum and minimum values of $R^{Guangzhou}$ and $R^{Shenzhen}$ reached the critical values of 1 and -1, respectively, indicating that the real estate markets in $R^{Guangzhou}$ and $R^{Shenzhen}$ fluctuate more violently than in $R^{Beijing}$ and $R^{Shanghai}$.

Table 3. The statistics of the RES index in four Chinese cities.

Metrics	Variance	Standard Deviation	Average	Median	Maximum	Minimum
$R^{Beijing}$	0.091	0.301	0.033	0.034	0.625	−0.731
$R^{Shanghai}$	0.010	0.315	0.076	0.074	0.750	−0.536
$R^{Guangzhou}$	0.111	0.334	0.116	0.112	1	−1
$R^{Shenzhen}$	0.228	0.478	0.063	0	1	−1

We infer that in cities with more dynamic economies, such as Guangzhou and Shenzhen, the real estate market is subject to rapid economic changes, which makes sentiment volatile and prone to extremes. Meanwhile, Beijing and Shanghai have relatively stable and larger real estate markets with less fluctuation in market sentiment.

Figure 4 shows the RES index R^{City} and the two kinds of real estate price index nP^{City} and sP^{City} from 2011 to 2022. Overall, the fluctuations of nP^{City} and sP^{City} in these four cities are complex, with no obvious cyclical trends, and only two distinct peaks can be observed. Specifically, Shenzhen has a high overall similarity among $nP^{Shenzhen}$ and $sP^{Shenzhen}$, while other cities have only local similarities. Furthermore, there is no apparent correlation trend between nP^{City}, sP^{City}, and R^{City}, and we need to construct VAR models for both the RES index and the real estate price index to analyze their latent correlations and predict future trends.

Figure 4. The RES index, R^{City}, and the two kinds of real estate price index, nP^{City} and sP^{City} from 2011 to 2022.

3.4. Construction of VAR Model for Real Estate Price Index

To explore the dynamic relationships between the RES index and the real estate price index, we employ Algorithm 2 to construct the two VAR models for each city. One is named nV^{city} and represents the VAR models for the RES index and the new housing price index, while the other is named sV^{city} and represents the VAR model for the RES index and the second-hand housing price index. Therefore, eight VAR models for four Chinese cities will be constructed to explore the relationships between the RES index and the new or second-hand housing price index.

We estimate the VAR model using the 10-year data series from 2011 to 2020 and predict and validate the outcomes for the years 2021 to 2022. We first perform an ADF stationarity test on these index series, then solve the optimal lag order based on the AIC criterion, and finally, construct the VAR model and predict the future RES index and real estate price index based on Algorithms 2 and 3.

Performance Matrices. To assess the predictive performance of nV^{city} and sV^{city}, we employ widely recognized evaluation metrics for regression analysis, i.e., Mean Absolute Error (MAE), Root Mean Square Error ($RMSE$), and R-Square (R^2). These indices are formally defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y} - y_i)^2}, \tag{13}$$

where $\hat{y}_i$ and y_i denote the predicted and true values of the i-th sample, respectively, and n is the number of samples. Smaller MAE and $RMSE$ indicate better forecasting performance. The value range of R^2 is [0,1], and the larger R^2 represents a better forecasting performance.

ADF stationary test. The prerequisite for constructing a VAR model is two or more stable data sequences. So, we first employ the ADF (Augmented Dickey–Fuller) testing method [60] to examine the stationarity of the new housing price index, the second-hand housing price index, and the RES index. We set the confidence level to 99%. If the significance test statistic is less than the critical value at the 1% significance level, we reject the null hypothesis of non-stationarity, implying a 99% confidence that the data series is stationary. For the two data sequences used to construct the VAR model, if either data series is found to be non-stationary, we apply Equation (5) to their difference series iteratively until both data series achieve stationarity.

Table 4 shows the ADF test results with a 99% confidence level. $I(0)$ represents the raw data series, and $I(1)$ represents a first-order differential data series. We can observe that the stationary first-order differential data series include $R^{Beijing}$, $R^{Shenzhen}$, $nP^{Guangzhou}$, $nP^{Shenzhen}$, and $sP^{Shanghai}$, and others are the stationary original data series. According to this, we can handle the two data series with the same differential order for constructing a VAR model.

Table 4. The ADF stationary test results with a 99% confidence level.

Cities	R^{City}	nP^{City}	sP^{City}
Beijing	$I(1)$	$I(0)$	$I(0)$
Shanghai	$I(0)$	$I(0)$	$I(1)$
Guangzhou	$I(0)$	$I(1)$	$I(0)$
Shenzhen	$I(1)$	$I(1)$	$I(0)$

Estimate the length of the lag period. We employ the AIC (Akaike Information Criterion) method [61] to estimate the lag length of the VAR model, which is the parameter k in Equation (6). AIC is a comprehensive evaluation metric that reflects both model complexity and fitting effect. A smaller AIC value indicates a better fitting effect and a moderate level of model complexity. Generally, we determine the lag length by selecting the minimum AIC value.

Figure 5 demonstrates the impact of the lag period length on AIC values. It is evident that all eight VAR models attain their minimum AIC values within 9 lag periods. Among

these, Shanghai's new housing price model exhibits the shortest lag period, which is 1, whereas Beijing's second-hand housing price model has the longest lag period, which is 8. This suggests that the VAR model for Shanghai's new housing price is relatively straightforward, as the current values of the variables in the model are influenced solely by data from one historical period. Conversely, the current values of the variables in Beijing's second-hand housing price model should consider data spanning the past 8 periods.

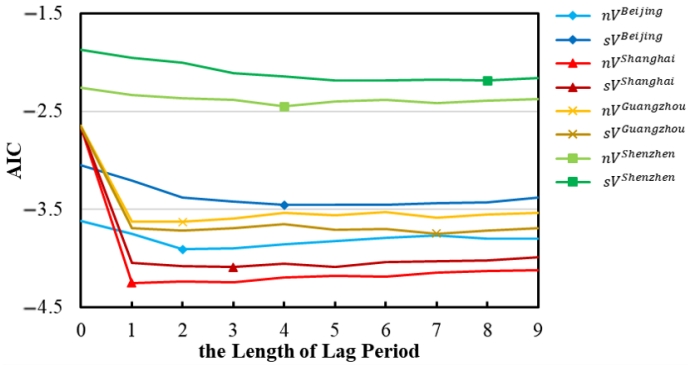

Figure 5. The effect of the lag period length on AIC values of the VAR model.

VAR modeling. After determining the optimal lag length, we can employ the OLS method to estimate the parameter matrix in the VAR model based on Equation (7), then construct two VAR models, i.e., the VAR model for the RES index and the new housing price index, nV^{City}, and the VAR model for the RES index and the second-hand housing price index, sV^{City}, for each city based on Equation (5).

To verify the performance of these constructed VAR models, we use data from 2011 to 2020 to train VAR models (the detailed parameters of the eight VAR models can be found in Supplementary Materials) and forecast the new and second-hand real estate price indices from 2021 to 2022. Table 5 presents the performance metrics of all eight VAR models. It is evident that all models exhibit good fitting performance, with the average MAE, RMSE, and R^2 being 0.128, 0.161, and 0.884, respectively. Furthermore, the RES index demonstrates a good fit with the new and second-hand housing price indices. This indicates a strong correlation between the RES index and these housing price indices.

Table 5. The performance of VAR models in housing price index forecasting.

Dataset	MAE	RMSE	R^2
$nV^{Beijing}$	**0.078**	**0.108**	0.842
$nV^{Shanghai}$	0.105	0.142	0.806
$nV^{Guangzhou}$	0.117	0.145	**0.941**
$nV^{Shenzhen}$	0.106	0.139	0.898
$sV^{Beijing}$	0.158	0.188	0.888
$sV^{Shanghai}$	0.177	0.214	0.862
$sV^{Guangzhou}$	0.160	0.202	0.905
$sV^{Shenzhen}$	**0.119**	**0.146**	**0.927**
Average	0.128	0.161	0.884

Among the four VAR models targeting the new housing price index, $nV^{Beijing}$ exhibits better performance, with the smallest MAE, 0.078, and RMSE, 0.108, while $nV^{Guangzhou}$ has the largest R^2 value of 0.941, indicating the best-fitting effect between the RES index and the new housing price index in Guangzhou. Additionally, $sV^{Shenzhen}$ outperforms the other VAR models targeting the second-hand housing price index, with the MAE, RMSE, and R^2 values of 0.119, 0.146, and 0.927, respectively.

Figure 6 presents the visualization of the forecasting results of the eight VAR models from 2021 to 2022. Obviously, the forecasting results of all eight models align well with the actual values, indicating that the VAR models can capture a dynamic correlation between the RES index and the housing price index. Furthermore, the fitting results of $nV^{Beijing}$ (Figure 6a) and $nV^{Guangzhou}$ (Figure 6e) are superior to those of $nV^{Shanghai}$ (Figure 6c) and $nV^{Shenzhen}$ (Figure 6g). Additionally, the fitting results of $sV^{Shenzhen}$ (Figure 6h) are better than those of the other three sV models (Figure 6b,d,f). This is consistent with the performance results presented in Table 5.

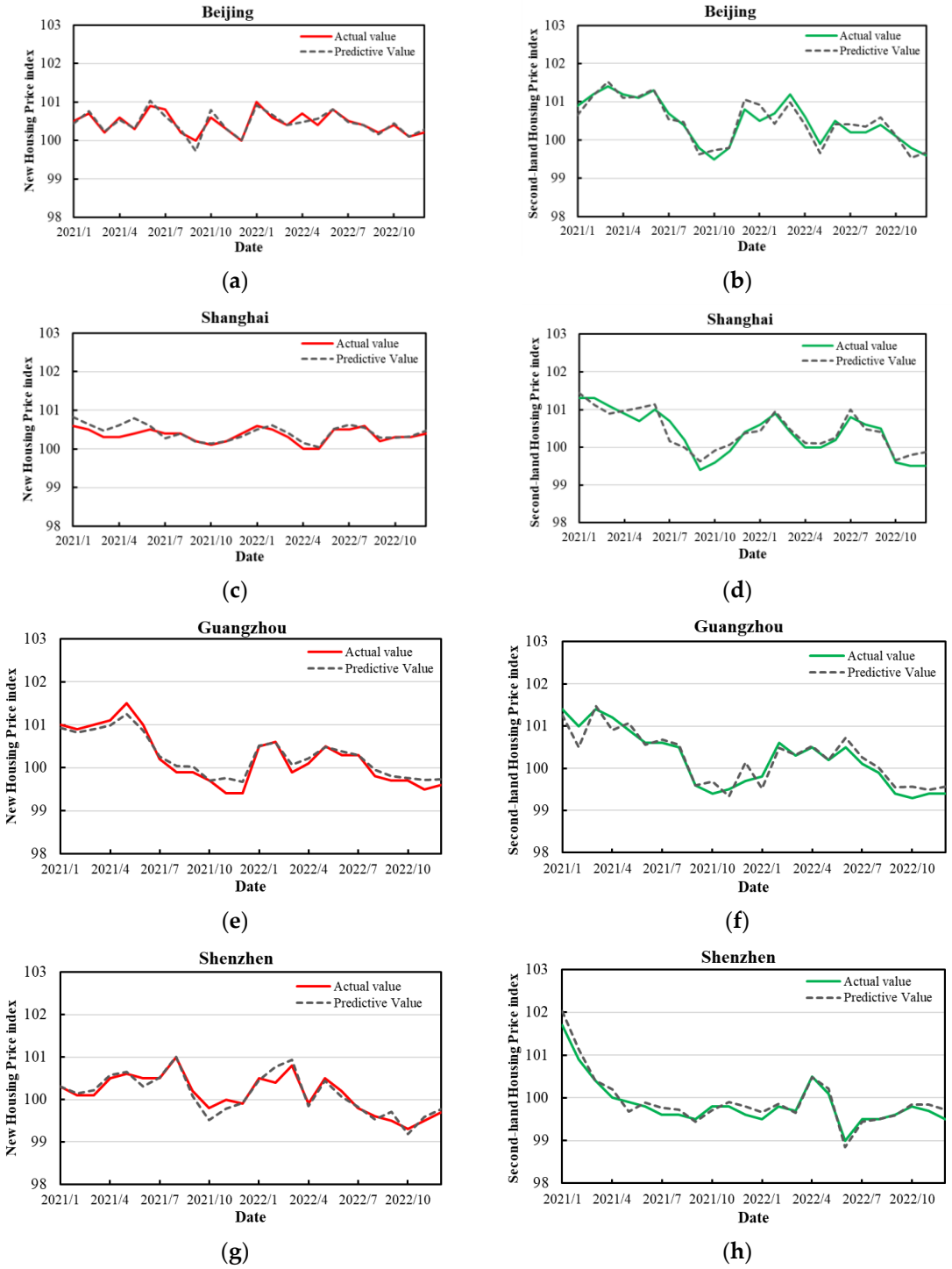

Figure 6. The forecasting results of all eight VAR models: (**a**) $nV^{Beijing}$, (**b**) $sV^{Beijing}$, (**c**) $nV^{Shanghai}$, (**d**) $sV^{Shanghai}$, (**e**) $nV^{Guangzhou}$, (**f**) $sV^{Guangzhou}$, (**g**) $nV^{Shenzhen}$, and (**h**) $sV^{Shenzhen}$.

3.5. Impulse Responses Analysis

To further analyze the dynamic relationships between the RES index and the new/second-hand housing price indices, we employ the Impulse Response Function (IRF) to quantify how the new/second-hand housing price index responds to impulses from the RES index.

Figure 7 illustrates the response of eight new/second-hand housing price indices to a unit pulse excitation applied to the RES index. We explore the impact of the RES index on these housing price indices from three aspects: response direction, response amplitude, and convergence period.

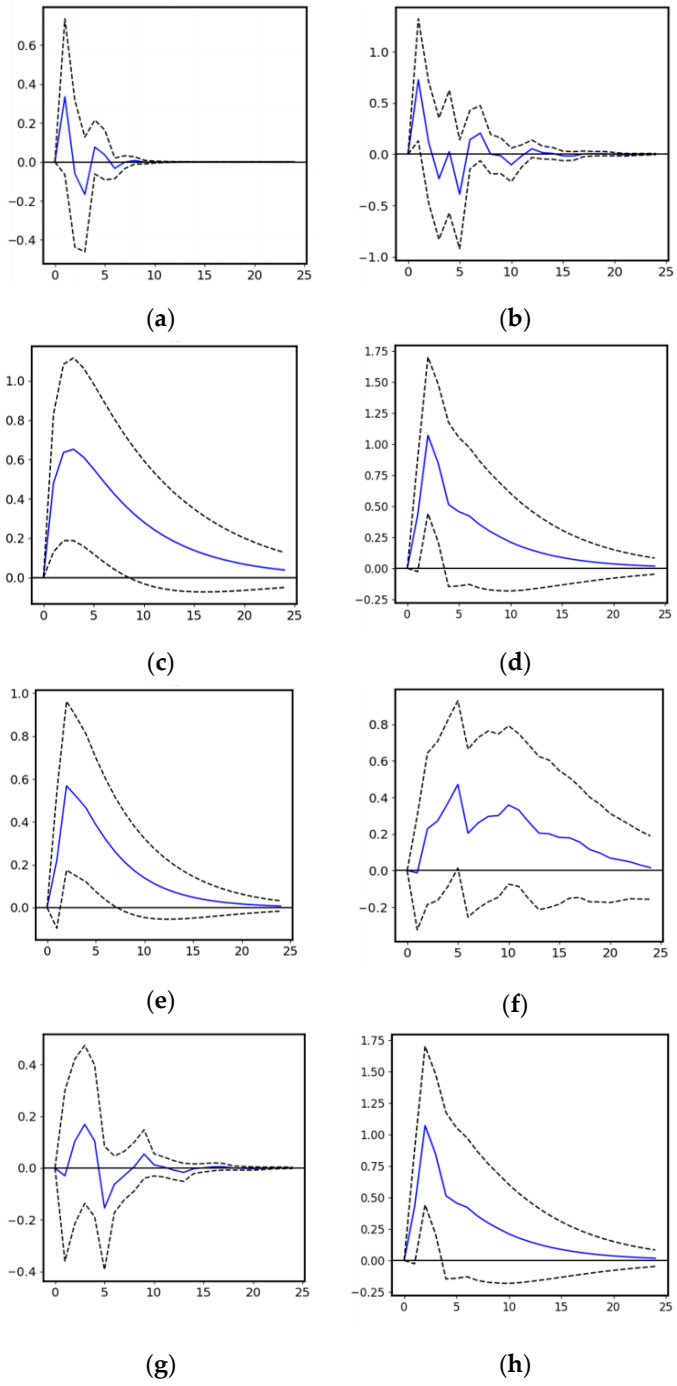

Figure 7. Impulse responses of eight price indices to the RES index: (**a**) $nV^{Beijing}$, (**b**) $sV^{Beijing}$, (**c**) $nV^{Shanghai}$, (**d**) $sV^{Shanghai}$, (**e**) $nV^{Guangzhou}$, (**f**) $sV^{Guangzhou}$, (**g**) $nV^{Shenzhen}$, and (**h**) $sV^{Shenzhen}$.

Response direction. The five models, $nV^{Shanghai}$, $nV^{Guangzhou}$, $sV^{Shanghai}$, $sV^{Guangzhou}$, and $sV^{Shenzhen}$, always provide the positive impulse responses of the RES index to the housing price index, indicating a positive correlation between them. This also suggests that the local real estate market in these cities is stable and market changes are in line with public expectations. Conversely, the other three models, $nV^{Beijing}$, $nV^{Shenzhen}$, and $sV^{Beijing}$, have chaotic response directions, implying that the local real estate market in these cities is constantly fluctuating and difficult to predict accurately.

Response amplitude. This metric reflects the sensitivity between news sentiment and the real estate market. Among the models, $\mathrm{sV}^{Shanghai}$, $\mathrm{sV}^{Guangzhou}$, and $\mathrm{sV}^{Shenzhen}$ exhibit the largest response amplitude, followed by $\mathrm{nV}^{Shangai}$, $\mathrm{nV}^{Guangzhou}$, and $\mathrm{sV}^{\mathrm{Beijing}}$, while $\mathrm{nV}^{\mathrm{Beijing}}$ and $\mathrm{nV}^{Shenzhen}$ showing the smallest values. Consequently, the RES indices for Shanghai, Guangzhou, and Shenzhen are the most sensitive to changes in the second-hand housing market.

Convergence period. This term represents the duration of the impact between news sentiment and the real estate market. Specifically, $nV^{Beijing}$ has the shortest convergence period, approximately 10 periods, while $nV^{Shenzhen}$ and $sV^{Beijing}$ have a convergence period of about 20, and other models have a convergence period of approximately 25.

Overall, the RES index of Beijing has the least impact on the housing price index, Shanghai and Guangzhou exhibit similar trends in both new and second-hand housing price indices, whereas Shenzhen displays a distinct difference between the new and second-hand housing price indices.

3.6. Summarizations

Based on the above results, the main findings are summarized below:

(1) Fine-tuning the BERT model on all real estate news datasets, the precision, recall, and F1 score consistently above 80%, indicating its good generalization ability for sentiment analysis tasks in Chinese real estate news datasets. Moreover, performance varied across cities, with $D^{Guangzhou}$ and $D^{Shenzhen}$ performing well on all metrics, while $D^{Beijing}$ performed relatively poorly, highlighting the influence of data quality and unique city characteristics on the model.

(2) The construction of China's urban RES index revealed interesting trends. The RES index series of each city exhibited chaos and disorder, lacking obvious periodicity. The statistical features of the RES index differed across cities, with Shenzhen showing more severe fluctuations, Guangzhou having a more optimistic market sentiment, and Beijing and Shanghai being relatively stable.

(3) The VAR model construction and analysis provided valuable insights. After conducting ADF stationary tests and estimating the lag period using the AIC method, eight VAR models were constructed for the four cities to explore the relationships between the RES index and the new/second-hand housing price index. These models exhibited good forecasting performance, with average MAE, RMSE, and R^2 values of 0.128, 0.161, and 0.884, respectively, indicating a strong correlation between the RES index and housing price indices.

(4) The impulse responses analysis further clarified the dynamic relationships between real estate price and online news sentiment. The response directions, amplitudes, and convergence periods of the new/second-hand housing price indices to impulses from the RES index varied across models and cities. Some cities like Shanghai, Guangzhou, and Shenzhen showed positive correlations and higher sensitivity, while others like Beijing had more chaotic response directions and lower sensitivity.

4. Conclusions

In this paper, we proposed a housing price index forecasting scheme based on massive online news data, aiming to achieve real-time forecasting of monthly real estate prices in Chinese cities. To capture the dynamic trends of public opinions on the real estate market, we developed a BERT-based model to analyze the sentiment of online news and defined the Real Estate Sentiment (RES) index as a representation of the monthly aggregated public sentiment derived from real estate online news. Furthermore, we constructed a VAR model incorporating the RES index and the official new/second-hand housing price indices to explore their interaction mechanism. Empirical research was conducted in Beijing, Shanghai, Guangzhou, and Shenzhen, China. The experimental results demonstrated that our proposed RES index has a strong quantifiable correlation with the two official housing price indices and can be utilized to accurately predict future housing price indices It not only offered a new perspective and method for understanding and evaluating trends in China's real estate market but also addressed the shortcomings of traditional forecasting methods and enhanced the accuracy of predicting future trends in the real estate market.

Furthermore, the results of this research provide certain references for the housing price regulation policies of the Chinese government. Firstly, in view of the role of market sentiment in China's housing prices, the Chinese government can attempt to regularly quantify and release the sentiment of the real estate market. This will help establish a rational understanding of the current housing market among participants and promote rational decision-making by home buyers and developers. Secondly, considering that the influence of market sentiment on housing prices in different cities is heterogeneous, when setting alarm values, it is necessary to adjust according to the actual situation in different cities. Thirdly, as some media may exaggerate the Chinese real estate market excessively, which leads to stimulating market sentiment, the Chinese government should guide media organizations to release objective and fair reports on the real estate market and avoid over-exaggerating housing prices at the same time, so as to reduce the information asymmetry in the real estate market.

Future work will focus on three aspects. (1) Broader investigation: We will validate the effectiveness of the proposed scheme in more different types of cities. (2) More accurate forecasting: We will develop the deep learning-based time series forecasting method to capture dynamic features precisely. (3) More reliable real estate sentiment index: We will incorporate more data sources and types to better capture the dynamics of the real estate market. (4) More diverse data: We will incorporate more data to more comprehensively and accurately reflect market dynamics.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/systems13010042/s1. The parameter details of VAR models in Beijing, Shanghai, Guangzhou, and Shenzhen, including fitting coefficients, residuals, and their statistical characteristics, are presented in a total of 24 Tables.

Author Contributions: T.X.: writing—original draft and editing, methodology, conceptualization, and funding acquisition. Y.Z.: writing—review and resources. J.Y.: writing—review and editing, supervision, and conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific and Technology Development Project of Henan Province China, grant number 242102210005.

Data Availability Statement: The datasets and codes used and analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hromada, E. Real Estate Insights on Mortgage Rates, Apartment Prices, and Rentals in Czech Republic. *Int. J. Econ. Sci.* **2024**, *13*, 13–29. [CrossRef]
2. Shao, J.; Yu, L.; Hong, J.; Wang, X. Forecasting house price index with social media sentiment: A decomposition–ensemble approach. *J. Forecast.* **2025**, *44*, 216–241. [CrossRef]
3. Wang, X.; Li, K.; Wu, J. House price index based on online listing information: The case of China. *J. Hous. Econ.* **2020**, *50*, 101715. [CrossRef]
4. Li, C.; Zhu, H.; Ye, X.; Jiang, C.; Dong, J.; Wang, D.; Wu, Y. Study on average housing prices in the inland capital cities of China by night-time light remote sensing and official statistics data. *Sci. Rep.* **2020**, *10*, 7732. [CrossRef]
5. Okuta, F.O.; Kivaa, T.; Kieti, R.; Okaka, J.O. Comparing simple and complex regression models in forecasting housing price: Case study from Kenya. *Int. J. Hous. Mark. Anal.* **2024**, *17*, 144–169. [CrossRef]
6. Rahman, S.; Masih, M. Increasing household debts and its relation to GDP, interest rate and house price: Malaysia's perspective. *Munich Pers. RePEc Arch.* **2014**, 62635.
7. Vaidynathan, D.; Kayal, P.; Maiti, M. Effects of economic factors on median list and selling prices in the US housing market. *Data Sci. Manag.* **2023**, *6*, 199–207. [CrossRef]
8. Bramley, G.; Watkins, D. Housebuilding, demographic change and affordability as outcomes of local planning decisions: Exploring interactions using a sub-regional model of housing markets in England. *Prog. Plan.* **2016**, *104*, 1–35. [CrossRef]
9. Case, K.E.; Shiller, R.J. Forecasting prices and excess returns in the housing market. *Real Estate Econ.* **1990**, *18*, 253–273. [CrossRef]
10. Rey-Blanco, D.; Zofío, J.L.; González-Arias, J. Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. *Expert Syst. Appl.* **2024**, *235*, 121059. [CrossRef]
11. Erol, I.; Unal, U. Internal migration and house prices in Australia. *Reg. Stud.* **2023**, *57*, 1207–1222. [CrossRef]
12. Hromada, E.; Bednar, O.; Pavelka, T. Real Estate Market at A Crossroad-Era of Affordable Housing Is Gone. *Int. J. Econ. Sci.* **2023**, *12*, 38–61. [CrossRef]
13. Chen, J.; Glaeser, E.; Wessel, D. JUE Insight: The (non-) effect of opportunity zones on housing prices. *J. Urban Econ.* **2023**, *133*, 103451. [CrossRef]
14. Venhoda, O. Application of DSTI and DTI macroprudential policy limits to the mortgage market in the Czech Republic for the year 2022. *Int. J. Econ. Sci.* **2022**, *11*, 105–116. [CrossRef]
15. Chen, S.-S.; Lin, T.-Y. Revisiting the link between house prices and monetary policy. *BE J. Macroecon.* **2022**, *22*, 481–515. [CrossRef]
16. Lukavec, M.; Čáp, V.; Čermáková, K. How permitting process length influences development costs and real estate prices. *Econ. Environ.* **2024**, *89*, 768. [CrossRef]
17. Wei, C.; Fu, M.; Wang, L.; Yang, H.; Tang, F.; Xiong, Y. The research development of hedonic price model-based real estate appraisal in the era of big data. *Land* **2022**, *11*, 334. [CrossRef]
18. Zaki, J.; Nayyar, A.; Dalal, S.; Ali, Z.H. House price prediction using hedonic pricing model and machine learning techniques. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7342. [CrossRef]
19. Abellana, A.; Devaraj, M. Hedonic Modeling for Predicting House Prices during Covid19 Pandemic in the Philippines. In Proceedings of the 2021 3rd International Conference on Management Science and Industrial Engineering, Osaka, Japan, 2–4 April 2021; pp. 21–26. [CrossRef]
20. Bogin, A.; Doerner, W.; Larson, W. Local house price dynamics: New indices and stylized facts. *Real Estate Econ.* **2019**, *47*, 365–398. [CrossRef]
21. Nagaraja, C.; Brown, L.; Wachter, S. Repeat sales house price index methodology. *J. Real Estate Lit.* **2014**, *22*, 23–46. [CrossRef]
22. Baroni, M.; Barthélémy, F.; Mokrane, M. Real estate prices: A Paris repeat sales residential index. *J. Real Estate Lit.* **2005**, *13*, 303–322. [CrossRef]
23. Pai, P.-F.; Wang, W.-C. Using machine learning models and actual transaction data for predicting real estate prices. *Appl. Sci.* **2020**, *10*, 5832. [CrossRef]
24. Mosig, J. The weighted averages algorithm revisited. *IEEE Trans. Antennas Propag.* **2012**, *60*, 2011–2018. [CrossRef]
25. Stulp, F.; Sigaud, O. Many regression algorithms, one unified model: A review. *Neural Netw.* **2015**, *69*, 60–79. [CrossRef]
26. Menke, W. Review of the generalized least squares method. *Surv. Geophys.* **2015**, *36*, 1–25. [CrossRef]
27. Zhan, W.; Hu, Y.; Zeng, W.; Fang, X.; Kang, X.; Li, D. Total Least Squares Estimation in Hedonic House Price Models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 159. [CrossRef]
28. Pan, J.-X.; Fang, K.-T.; Pan, J.-X.; Fang, K.-T. Maximum likelihood estimation. Growth curve models and statistical diagnostics. In *Growth Curve Models and Statistical Diagnostics*; Springer Nature: London, UK, 2002; pp. 77–158. [CrossRef]
29. Maguire, P.; Miller, R.; Moser, P.; Maguire, R. A robust house price index using sparse and frugal data. *J. Prop. Res.* **2016**, *33*, 293–308. [CrossRef]
30. Goh, Y.M.; Costello, G.; Schwann, G. Accuracy and robustness of house price index methods. *Hous. Stud.* **2012**, *27*, 643–666. [CrossRef]

31. Wang, J.; Athanasopoulos, G.; Hyndman, R.J.; Wang, S. Crude oil price forecasting based on internet concern using an extreme learning machine. *Int. J. Forecast.* **2018**, *34*, 665–677. [CrossRef]
32. Coble, D.; Pincheira, P. Forecasting building permits with Google Trends. *Empir. Econ.* **2021**, *61*, 3315–3345. [CrossRef]
33. Gunter, B.; Koteyko, N.; Atanasova, D. Sentiment analysis: A market-relevant and reliable measure of public feeling? *Int. J. Mark. Res.* **2014**, *56*, 231–247. [CrossRef]
34. Fraiberger, S.P. News sentiment and cross-country fluctuations. *SSRN Soc. Sci. Resarch Netw.* **2016**, 125–131. [CrossRef]
35. Anese, G.; Corazza, M.; Costola, M.; Pelizzon, L. Impact of public news sentiment on stock market index return and volatility. *Comput. Manag. Sci.* **2023**, *20*, 20. [CrossRef]
36. Kavitha, G.; Saveen, B.; Imtiaz, N. Discovering public opinions by performing sentimental analysis on real time Twitter data. In Proceedings of the 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, 21–22 December 2018; pp. 1–4. [CrossRef]
37. Kang, J.; Lee, H.J.; Jeong, S.H.; Lee, H.S.; Oh, K.J. Developing a forecasting model for real estate auction prices using artificial intelligence. *Sustainability* **2020**, *12*, 2899. [CrossRef]
38. Nowak, A.D.; Price, B.S.; Smith, P.S. Real estate dictionaries across space and time. *J. Real Estate Financ. Econ.* **2021**, *62*, 139–163. [CrossRef]
39. Isler, O.; Flew, T.; Erol, I.; Dulleck, U. Market news and credibility cues improve house price predictions: An experiment on bounded rationality in real estate. *J. Behav. Exp. Financ.* **2021**, *31*, 100550. [CrossRef]
40. Hausler, J.; Ruscheinsky, J.; Lang, M. News-based sentiment analysis in real estate: A machine learning approach. *J. Prop. Res.* **2018**, *35*, 344–371. [CrossRef]
41. Soo, C.K. Quantifying sentiment with news media across local housing markets. *Rev. Financ. Stud.* **2018**, *31*, 3689–3719. [CrossRef]
42. Beracha, E.; Lang, M.; Hausler, J. On the relationship between market sentiment and commercial real estate performance—A textual analysis examination. *J. Real Estate Res.* **2019**, *41*, 605–638. [CrossRef]
43. McCarthy, S.; Alaghband, G. Enhancing financial market analysis and prediction with emotion corpora and news co-occurrence network. *J. Risk Financ. Manag.* **2023**, *16*, 226. [CrossRef]
44. Kim, S.; Kwon, M.J.; Kim, H.H. Sentiment Analysis of News Based on Generative AI and Real Estate Price Prediction: Application of LSTM and VAR Models. *Trans. Korea Inf. Process. Soc.* **2024**, *13*, 209–216. [CrossRef]
45. Shao, J.; Hong, J.; Wang, X.; Yan, X. The relationship between social media sentiment and house prices in China: Evidence from text mining and wavelet analysis. *Financ. Res. Lett.* **2023**, *57*, 104212. [CrossRef]
46. Calainho, F.D.; van de Minne, A.M.; Francke, M.K. A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate. *J. Real Estate Financ. Econ.* **2024**, *68*, 624–653. [CrossRef]
47. Ben Ameur, H.; Boubaker, S.; Ftiti, Z.; Louhichi, W.; Tissaoui, K. Forecasting commodity prices: Empirical evidence using deep learning tools. *Ann. Oper. Res.* **2024**, *339*, 349–367. [CrossRef]
48. Xiao, H.; Luo, L. An Automatic Sentiment Analysis Method for Short Texts Based on Transformer-BERT Hybrid Model. *IEEE Access* **2024**, *12*, 93305–93317. [CrossRef]
49. Adelakun, N.O.; Baale, A.A. Sentiment analysis of financial news using the bert model. *ITEGAM-JETIA* **2024**, *10*, 21–27. [CrossRef]
50. Wen, Y.; Liang, Y.; Zhu, X. Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China. *PLoS ONE* **2023**, *18*, e0275382. [CrossRef]
51. Kazi, A.; Kumar, G.; Agrawal, R. Leveraging Bidirectional Encoder Representations from Transformers (BERT) for Enhanced Sentiment Analysis. In *Advances in Computational Intelligence and Informatics, Proceedings of the International Conference on Advances in Computational Intelligence and Informatics, Hyderabad, India, 22–23 December 2023*; Springer Nature: London, UK, 2024; pp. 87–95. [CrossRef]
52. Cao, Y.; Sun, Z.; Li, L.; Mo, W. A study of sentiment analysis algorithms for agricultural product reviews based on improved bert model. *Symmetry* **2022**, *14*, 1604. [CrossRef]
53. Piispanen, A. Price Determinants of Airbnb Apartments: An Approach with Deep Language Representations. Master's Thesis, Tampere University, Tampere, Finland, 2021; p. 102709. [CrossRef]
54. Yang, H.; Ye, C.; Lin, X.; Zhou, H. Stock Market Prediction Based on BERT Embedding and News Sentiment Analysis. In *Service Science, Proceedings of the International Conference on Service Science, Harbin, China, 13–14 May 2023*; Springer Nature: London, UK, 2023; pp. 279–291. [CrossRef]
55. Haley, C. This is a BERT. Now there are several of them. Can they generalize to novel words? In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Online, 20 November 2020; pp. 333–341. [CrossRef]
56. Stock, J.H.; Watson, M.W. Vector autoregressions. *J. Econ. Perspect.* **2001**, *15*, 101–115. [CrossRef]
57. Sims, C.A. Vector Autoregressions and Reality: Comment. *J. Bus. Econ. Stat.* **1987**, *5*, 443–449. [CrossRef]
58. Wu, Y. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.

59. Mannor, S.; Peleg, D.; Rubinstein, R. The cross entropy method for classification. In Proceedings of the 22nd international conference on Machine learning, Bonn, Germany, 7–11 August 2005; pp. 561–568. [CrossRef]
60. Dickey, D.A.; Fuller, W.A. Likelihood ratio statistics for autoregressive time series with a unit root. *Econom. J. Econom. Soc.* **1981**, *49*, 1057–1072. [CrossRef]
61. Akaike, H. Factor analysis and AIC. *Psychometrika* **1987**, *52*, 317–332. [CrossRef]
62. Sharmaasasdad, S.C. The effects of autocorrelation among errors on the consistency property of OLS variance estimator. *J. Econom.* **1985**, *27*, 335–361. [CrossRef]
63. Lee, L.-F. Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econom. Theory* **2002**, *18*, 252–277. [CrossRef]
64. Diederik, P.K. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]